PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-135804
(43)Date of publication of application : 10.05.2002

(51)Int.Cl. H04N 11/04
G06T 7/20
H04N 5/92
H04N 7/24

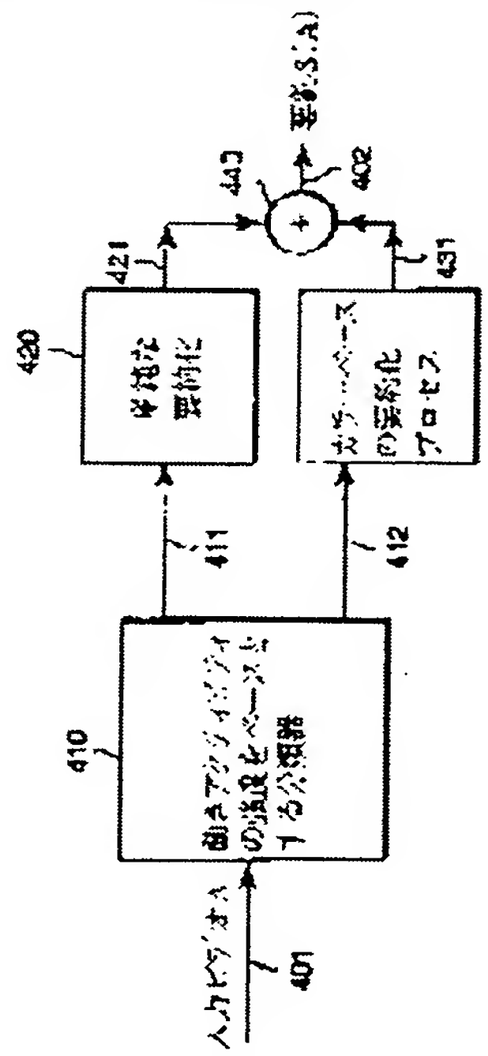
(21)Application number : 2001-229656
(22)Date of filing : 30.07.2001

(71)Applicant : MITSUBISHI ELECTRIC RESEARCH LABORATORIES INC
(72)Inventor : DIVAKARAN AJAY
PEKER KADIR A
SUN HUIFANG

(30)Priority
Priority number : 2000 634364 Priority date : 09.08.2000 Priority country : US

(54) METHOD FOR SUMMARIZING VIDEO BY USING MOTION DESCRIPTOR AND COLOR DESCRIPTOR

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a method for summarizing compressed video.
SOLUTION: In this method, the intensity of a motion activity is extracted from a shot in the compressed video. Next, a segment easy to summarize the video and a segment difficult to summarize the video are divided by using the intensity of the motion activity. The segment easy to summarize the video is expressed with an arbitrary frame selected out of the segment easy to summarize the video and on the other hand, a frame sequence is generated from each of segments difficult to summarize the video by the summarization process of color base. By combining the selected frame and generated frame in each of segments in each of shots, the summary of the compressed video is formed.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2002－135804
(P2002－135804A)
(43)公開日　平成14年５月10日(2002.5.10)

(51)Int.Cl.'	識別記号	FI		テーマコード'(参考)
H04N 11/04		H04N 11/04	Z	5C053
G06T 7/20		G06T 7/20	C	5C057
H04N 5/92		H04N 7/13	Z	5C059
7/24		5/92	H	5L096

審査請求　未請求　請求項の数８　ＯＬ　外国語出願　（全 31 頁）

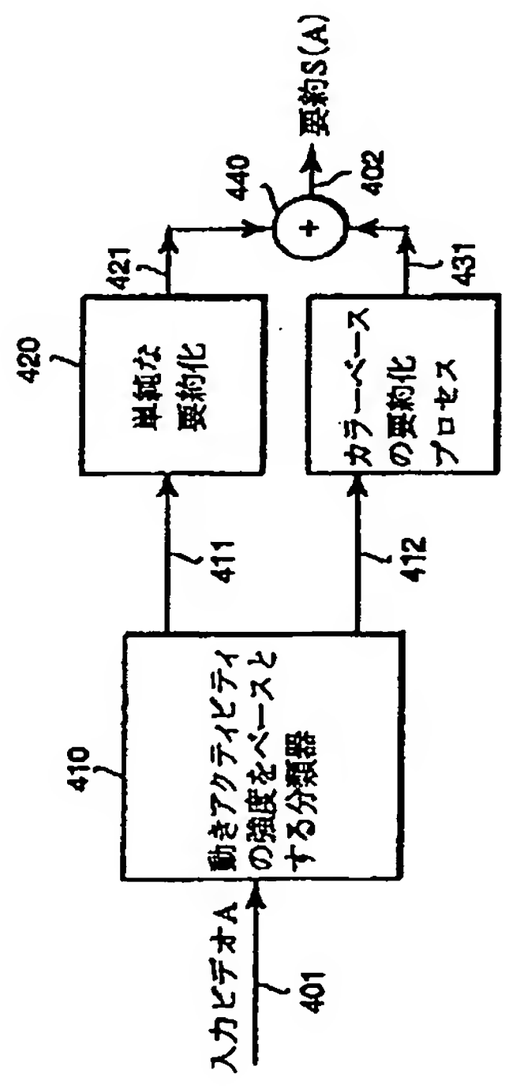
(21)出願番号　特願2001－229656(P2001－229656)
(22)出願日　平成13年７月30日(2001.7.30)
(31)優先権主張番号　０９／６３４３６４
(32)優先日　平成12年８月９日(2000.8.9)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　597067574
ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
アメリカ合衆国、マサチューセッツ州、ケンブリッジ、ブロードウエイ　201
(72)発明者　アジェイ・ディヴァカラン
アメリカ合衆国、ニュージャージー州、デンヴィル、コブ・ロード　20
(74)代理人　100057874
弁理士　曾我　道照　（外６名）

最終頁に続く

(54)【発明の名称】　動き記述子およびカラー記述子を用いてビデオを要約化する方法

(57)【要約】　（修正有）
【課題】圧縮ビデオを要約する方法を提供する。
【解決手段】方法は、圧縮ビデオにおけるショットから動きアクティビティの強度を抽出する。該方法は次に、動きアクティビティの強度を用いて、ビデオを要約化が容易なセグメントと困難なセグメントとに区分化する。要約化が容易なセグメントは、要約化が容易なセグメントから選択される任意のフレームで表される一方、カラーベースの要約化プロセスが要約化が困難なセグメントそれぞれからフレームシーケンスを生成する。各ショットにおける各セグメントの選択されたフレームおよび生成されたフレームを組み合わせて、圧縮ビデオの要約を形成する。



【特許請求の範囲】
【請求項１】　動き特徴およびカラー特徴を含む圧縮ビデオを要約化する方法であって、
前記圧縮ビデオを複数のショットに区分化するステップと、
各ショットの各フレームを前記動き特徴に従って、比較的低い動きアクティビティを有する第１のクラスのフレームと、比較的高い動きアクティビティを有する第２のクラスのフレームとに分類するステップと、
前記同じ分類を有する連続フレームをセグメントにグループ化するステップと、
前記第１のクラスの各セグメントから任意の１つまたは複数のフレームを選択するステップと、
前記カラー特徴を用いて、前記第２のクラスの各セグメントからフレームのシーケンスを生成するステップと、
各ショットの各セグメントの前記選択されたフレームおよび前記生成されたフレームを組み合わせて、前記圧縮ビデオの要約を形成するステップと、を含む、方法。
【請求項２】　前記選択されたフレームおよび前記生成されたフレームを時間的順序で組み合わせるステップをさらに含む、請求項１記載の方法。
【請求項３】　前記選択されたフレームおよび前記生成されたフレームを空間的順序で組み合わせるステップをさらに含む、請求項１記載の方法。
【請求項４】　前記選択されたフレームおよび前記生成されたフレームのサイズを縮小して縮小フレームを形成するステップをさらに含む、請求項３記載の方法。
【請求項５】　前記選択されたフレームおよび前記生成されたフレームを意味論的順序で組み合わせるステップをさらに含む、請求項１記載の方法。
【請求項６】　前記第２のクラスの各セグメントのフレームを、前記カラー特徴に従ってクラスタにグループ化するステップと、
各クラスタにクラスタの要約を生成するステップと、
前記クラスタの要約を組み合わせて、前記生成されたフレームのシーケンスを形成するステップと、をさらに含む、請求項１記載の方法。
【請求項７】　前記要約は、前記ビデオの再生中に作成される、請求項１記載の方法。
【請求項８】　前記要約は、前記ビデオの再要約化に使用される、請求項１記載の方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明はビデオに関し、特に、圧縮ビデオの要約化に関する。
【０００２】
【従来の技術】ビデオの要約を自動的に生成すること、特に圧縮デジタルビデオから要約を生成することが望ましい。
【０００３】圧縮ビデオフォーマット

ビデオをデジタル信号として圧縮する基本的な方法が、ＭＰＥＧ（Motion Picture Expert Group）に採用されている。ＭＰＥＧ規格は、画像のフルフレームについての情報を時折作成することで、高いデータ圧縮率を達成する。フル画像フレーム、すなわちフレーム内符号化フレームは「Ｉフレーム」または「アンカーフレーム」と呼ばれることが多く、あらゆる他のフレームとは独立したフルフレーム情報を含む。画像差フレーム、すなわちフレーム間符号化フレームは、「Ｂフレーム」および「Ｐフレーム」、または「予測フレーム」と呼ばれることが多く、これは、Ｉフレーム間で符号化され、基準フレームからの画像差、すなわち残余のみを反映する。
【０００４】通常、ビデオシーケンスの各フレームは、より小さな画素、すなわちピクセルのデータブロックに分割される。各ブロックは離散コサイン変換（ＤＣＴ）関数を施され、統計上依存した空間領域ピクセルを独立周波数領域ＤＣＴ係数に変換する。「マクロブロック」と呼ばれる、それぞれの８ｘ８、または１６ｘ１６ブロックのピクセルは、ＤＣＴ関数を施されて、符号化信号を提供する。
【０００５】ＤＣＴ係数は通常エネルギー集中的であるため、１つのマクロブロックにおいて少数の係数のみが、ピクチャ情報の主要部分を含む。たとえば、マクロブロックがオブジェクトのエッジ境界を含む場合、変換後のそのブロックのエネルギー、すなわちＤＣＴ係数で表されるものは比較的大きなＤＣ係数を含み、係数のマトリクスにわたってＡＣ係数がランダムに分散される。
【０００６】一方、非エッジマクロブロックは通常、同様の大きなＤＣ係数と、そのブロックに関連する他の係数よりも実質的に大きな隣接ＡＣ係数とを特徴とする。ＤＣＴ係数は通常、適応量子化を施されてから、送信媒体に関してランレングス符号化および可変長符号化される。このため、送信データのマクロブロックは通常、８ｘ８のマトリクスに満たない符号語を含む。
【０００７】フレーム内符号化データ、すなわち符号化ＰまたはＢフレームデータのマクロブロックは、予測ピクセルとマクロブロックにおける実際のピクセルの間の差分のみを表す。フレーム内符号化データおよびフレーム間符号化データのマクロブロックはまた、用いられた量子化のレベル、マクロブロックのアドレスインジケータまたはロケーションインジケータ、ならびにマクロブロックのタイプ等の情報も含む。後者の情報は、「ヘッダ」または「オーバーヘッド」情報と呼ばれる。
【０００８】各Ｐフレームは、最後のＩフレームまたはＰフレームから予測される。各Ｂフレームは、これを挟むＩフレームまたはＰフレームから予測される。予測符号化プロセスは、Ｉフレームのどのマクロブロックの変位量が現在符号化されているＢフレームまたはＰフレームのマクロブロックと最も密接にマッチしているかを示す、「動きベクトル」としばしば呼ばれる変位ベクトル

の生成を含む。Ｉフレームにおけるマッチするブロックのピクセルデータが、符号化されているＰフレームまたはＢフレームのブロックからピクセル毎に減じられ、残余が現れる。変換された残余およびベクトルは、Ｐフレームまたはフレームの符号化データの一部を形成する。

【０００９】ビデオ分析

ビデオ分析は、ビデオコンテンツの理解を意図してのビデオ処理として定義することができる。ビデオの理解は、「低レベル」の統辞語的理解から「高レベル」の意味論的理解までの範囲としうる。

【００１０】低レベルの理解は、カラー、モーション、テクスチャ、形状等、低レベルの特徴を分析することでなされる。低レベルの特徴は、ビデオを「ショット」に区分化することのために用いることができる。本明細書において、ショットは、カメラへの電源投入時に始まり、カメラの電源オフまで続くフレームシーケンスとして定義される。通常、ショットにおけるフレームシーケンスは、単一の「シーン」を取り込む。低レベルの特徴を用いて、記述子を生成することが可能である。そして、記述子を用いて、たとえばビデオにおける各ショットおよびおそらくその長さの索引など、ビデオを索引付けることができる。

【００１１】ビデオの意味論的理解は、コンテンツの部類に関係し、その統語論的な構造には関係ない。たとえば、高レベルの特徴は、ビデオがアクションビデオであるか、ミュージックビデオであるか、「トーキングヘッド（talking head：画面に話し手が登場するもの）」のビデオであるかなどを表す。

【００１２】ビデオ要約化

ビデオ要約化とは、ビデオの意味論的本質を伝えるビデオのコンパクトな表現を作成することと定義することができる。コンパクトな表現には、「キー」フレームまたは「キー」セグメント、あるいはキーフレームとセグメントの組み合わせを含めることができる。一例として、テニスの試合のビデオ要約は、２つのフレーム、すなわち双方の選手を取り込んだ第１のフレームと、トロフィーを持った勝者を取り込んだ第２のフレームと、を含みうる。より詳細かつ長い要約には、マッチポイントを取り込んだすべてのフレームをさらに含めることができる。このような要約を手動で生成することは確かに可能であるが、これには時間と費用がかかる。したがって、自動要約化が望まれる。

【００１３】自動ビデオ要約化方法は周知であり、S. Pfeifer他による「Abstracting Digital Movies Automatically」（J. Visual Comm. Image Representation, vol. 7, no. 4, pp. 345 － 353, １９９６年１２月）および「An Integrated Schemefor Automated Video Abstraction Based on Unsupervised Cluster － Validity Analysis」（IEEE Trans. On Circuits and Systems for Video Technology,Vol. 9, No. 8, １９９９年１２月）を参照されたい。

【００１４】最も良く知られているビデオ要約化方法は、専らカラーベースの要約化に集中している。Pfeiffer他だけが、ビデオ要約の作成に、動きを他の特徴と組み合わせて用いている。しかし、Pfeiffer他による方法は単に、組み合わせた特徴間で考えられうる修正を見過ごした重み付き組み合わせを用いるだけにすぎない。要約化方法によっては、動き特徴を用いてキーフレームを抽出するものもある。

【００１５】図１に示すように、従来技術によるビデオ要約化方法は殆ど、カラー特徴に基づくクラスタ化を強調している。これは、カラー特徴が抽出し易くかつノイズに耐性があるためである。典型的な方法は、ビデオＡ１０１を入力としてとり、カラーベースの要約化プロセス１００を適用してビデオ要約Ｓ（Ａ）１０２を作成する。ビデオ要約は、ビデオ全体の１つの要約、あるいは関心を引くフレームのセットのいずれかからなる。

【００１６】方法１００は通常、以下のステップを含む。第１に、カラー特徴に従いビデオのフレームをクラスタ化する。第２に、クラスタをアクセスし易い階層データ構造にする。第３に、各クラスタからキーフレームまたはキーフレームシーケンスを抽出して、要約化する。

【００１７】動きアクティビティ記述子

ビデオはまた、様々なレベルのアクティビティ、またはアクション強度を有するものとして直観的に理解することができる。比較的高レベルのアクティビティの例は、スポーツイベントビデオでの得点チャンスであり、一方ニュースビデオは比較的低レベルのアクティビティを有する。最近提案されたＭＰＥＧ－７映像規格は、ビデオでの動きアクティビティに関連する記述子を提供する。

【００１８】

【発明が解決しようとする課題】本発明の目的は、カラーおよびテクスチャの特徴等、他の低レベルの特徴と組み合わせて、動き特徴、特に動きアクティビティ特徴を用いて自動ビデオ要約化方法を提供することである。

【００１９】

【課題を解決するための手段】本発明の背後にある主な洞察は、次の仮定を根拠としている。ビデオの動きアクティビティは、ビデオの要約化の相対的な困難さを示すのによい目安である。動きの量が多くなるほど、そのビデオの要約化は困難である。ビデオ要約は、包含するフレームの数、たとえばキーフレームの数、またはキーセグメントのフレームの数等により、量的に記述することができる。

【００２０】ビデオの動きアクティビティの相対強度は、カラー特徴の変化と強く相関する。換言すれば、動きアクティビティの強度が高い場合、カラー特徴の変化も高い可能性が高い。カラー特徴の変化が高い場合、カ

ラー特徴をベースとする要約には比較的多数のフレームが含まれることになり、カラー特徴の変化が低い場合には、要約にはより少数のフレームが含まれることになる。

【００２１】たとえば、「トーキングヘッド」ビデオでは、動きアクティビティのレベルが低く、また同様にカラー変化もごくわずかである。要約化がキーフレームをベースとする場合、ビデオの要約化には１つのキーフレームで十分である。キーセグメントを用いる場合、視覚的なビデオの要約化には、１秒のフレームシーケンスで十分である。一方、スポーツイベントでの得点チャンスでは動きアクティビティ強度およびカラー変化が非常に高く、したがって要約化には数個のキーフレームまたは数秒が必要である。

【００２２】より具体的に、本発明は、まずビデオから動きアクティビティの強度を抽出することで、ビデオの要約化をする方法を提供する。次に、該動きアクティビティの強度を用いて、ビデオを要約化の容易なセグメントと困難なセグメントとに区分する。

【００２３】要約化が容易なセグメントは、単一フレーム、またはセグメントのどこかから選択されたフレームで表される。要約化が容易なセグメントではフレーム間の差がごくわずかであるため、このフレームはいずれのフレームでもよい。困難なセグメントの要約化には、カラーベースの要約化プロセスを用いる。このプロセスは、要約化の困難な各セグメントからフレームシーケンスを抽出する。単一フレームおよび抽出したフレームシーケンスを組み合わせて、ビデオの要約を形成する。

【００２４】組み合わせには、時間的、空間的、または意味論的な順序付けを用いることができる。時間的配置の場合、たとえば最初から最後に、または最後から最初になどある時間順序で、フレームを繋ぎ合わせる。空間的配置の場合、フレームの縮小版を組み合わせてモザイクやある配列、たとえば長方形にし、１つのフレームで選択された要約フレームのいくつかの縮小版を見せる。意味論的順序の要約は、興奮度の高いものから低いものに移行する。

【００２５】

【発明の実施の形態】本発明は、カラー特徴および動き特徴を用いて、圧縮ビデオを要約化する。したがって、本発明の要約化方法は最初に、圧縮ビデオから特徴を抽出する。

【００２６】特徴抽出

カラー特徴

既知の技術を用いて、Ｉフレームのｄｃ係数を正確かつ容易に抽出することができる。ＰおよびＢのフレームの場合、ＤＣ係数は、完全に圧縮解除することなく動きベクトルを用いて近似することができる。たとえば、Ｙｅｏ他著「On theExtraction of DC Sequence from MPEG video」（IEEE ICIP Vol. 2, １９９５年）を参照されたい。ＤＣ画像のＹＵＶ値は、カラー特徴を抽出するために、別の色空間に変形することができる。

【００２７】最も一般に使用される技術は、カラーヒストグラムを用いる。カラーヒストグラムは、画像および映像の索引付けおよび検索に広く用いられてきている。Smith他による「Automated Image Retrieval Using Color and Texture」（IEEE Transaction on Pattern Analysis and Machine Intelligence, １９９６年１１月）を参照されたい。通常、３チャネルＲＧＢ色空間では、各チャネルに４つのビンがあり、カラーヒストグラムには総計６４（４×４×４）個のビンが必要である。

【００２８】動き特徴

動き情報は、大抵動きベクトルに埋め込まれる。動きベクトルは、ＰフレームおよびＢフレームから抽出することができる。動きベクトルは通常、実際の光の流れに対する荒く、散在した近似であるため、ここでは動きベクトルを質的にのみ用いる。動きベクトルを用いる多くの異なる方法が知られている。Tan他による「A new method for camera motion parameter estimation」（Proc. IEEE International Conference on Image Processing, Vol. 2, pp. 722 － 726, １９９５年）、IEEE Trans. on Circuits and Systems for Video Technology １９９９年に見られるTan他による「Rapid estimation of camera motion from compressed video with application to video annotation」、Kobla他による「Detection of slow-motion replay sequences for identifying sports videos」（Proc. IEEE Workshop on Multimedia Signal Processing, １９９９年）、Kobla他による「Special effect edit detection using VideoTrails: a comparisonwith existing techniques」（Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases VII, １９９９年）、Kobla他による「Compressed domain video indexing techniques using DCT and motion vector information in MPEG video」（Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases V, SPIE Vol. 3022, pp. 200 － 211, １９９７年）、およびMeng他による「CVEPS － a compressed video editing and parsing system」（Proc. ACM Multimedia 96, １９９６年）を参照されたい。

【００２９】上述したように、殆どの従来技術による要約化方法は、カラー特徴のクラスタ化をベースとして、カラー記述子を得る。カラー記述子は、定義上比較的ノイズに耐性があるが、ビデオの動き特徴を含まない。しかし、動き記述子はノイズに対する耐性がより低い傾向があるため、ビデオの要約化に広くは用いられていない。

【００３０】Divakaran他出願の米国特許出願第０９／４０６，４４４号「Activity Descriptor for Video Sequences」は、圧縮画像における動きベクトルから導出

した動き特徴を用いて、ビデオにおける動きアクティビティおよびビデオにおける動きアクティビティの空間分布を決定する方法を記載している。このような記述子は、ビデオ閲覧への適用に関してはうまくいく。本明細書では、このような動き記述子をビデオ要約化に適用する。

【００３１】本明細書では、ビデオにおけるアクティビティの相対レベルを用いて、ビデオの「要約化難易度」を測定できるものと仮定する。不都合なことに、この仮定をテストする単純で客観的な尺度はない。しかし、動きの変化はカラー特徴の変化によってなされることが多いため、動きアクティビティの相対強度とビデオのカラー特徴変化の間の関係を調べる。

【００３２】動きおよび色の変化

ＭＰＥＧ－７「テストセット」から、ビデオのカラー特徴および動き特徴を抽出することで、これを行う。動きベクトルの大きさの平均を計算することですべてのＰフレームから動きアクティビティ特徴を抽出するとともに、すべてのＩフレームから６４ビンＲＧＢヒストグラムを抽出する。次に、ＩフレームからＩフレームへのヒストグラムにおける変化を計算する。メジアンフィルタをフレーム毎のカラーヒストグラム変化のベクトルに適用し、セグメントカット（segmentcut）または他のセグメント遷移に対応する変化をなくす。図２および図３に示すように、フレーム毎に動きアクティビティ強度対メジアンフィルタ済カラー変化をグラフ化する。

【００３３】図２および図３はそれぞれ、「jornaldanoitel」および「news1」テストセットについての動きアクティビティ強度とカラー相違の間の関係を示す。動きアクティビティ強度とカラー変化の間には明確な相関がある。アクティビティが低い場合、カラー変化も低いことは非常にはっきりとしている。アクティビティレベルが高いほど、高アクティビティの予想されるソースが多くなり、カラーコンテンツ変化が原因ではないものもありうるため、相関は明確ではなくなる。しかし、アクティビティが非常に低い場合、コンテンツがフレーム毎に変化しない可能性が高い。本発明では、この情報を用いてビデオを予めフィルタリングし、殆ど静的なセグメントを検出し、したがって静的セグメントを単一のキーフレームで要約化する。これらの結果に基づき、以下の要約化方法を提供する。

【００３４】要約化方法

図４は、入力圧縮ビデオＡ４０１を要約化して要約Ｓ（Ａ）４０２を作成するための方法４００を示す。

【００３５】入力圧縮ビデオ４０１は、当分野で周知であり、かつ上述した標準的な技術を用いて「ショット」に区分化される。最初にビデオをショットに区分化することで、各ショットが同質であり、シーン変更を含まないよう確実にする。よって、たとえば、意味論的レベルでは同一に見える連続した異なる１０の「トーキングヘッド」ショットのビデオを適宜要約化する。この点から、ビデオはショット毎に処理される。

【００３６】ステップ４１０は、各ショットの各フレームについての動きアクティビティの相対強度を決定する。各フレームは、第１あるいは第２のクラスのいずれかに分類される。第１のクラスは、要約化が比較的容易なフレームを包含し、第２のクラス４１２は要約化が比較的困難なフレームを包含する。換言すれば、この分類は動きをベースとするものである。

【００３７】分類が同じである各ショットの連続フレームは、要約化が「容易」なセグメント４１１あるいは要約化が「困難」なセグメント４１２のいずれかにグループ化される。

【００３８】各ショットの容易なセグメント４１１では、セグメントからキーフレームまたはキーフレームシーケンスを選択する（４２１）ことで、セグメントの単純な要約化４２０を行う。容易なセグメント中のフレームはすべて意味論的に同様であるものとみなされるため、選択されるキーフレーム４２１は、セグメント中の任意のフレームであることができる。

【００３９】各ショットの困難なセグメント４１２については、カラーベースの要約化プロセス５００を適用して、セグメントをキーフレームシーケンス４３１として要約化する。

【００４０】各ショットのキーフレーム４２１および４３１を組み合わせて各ショットの要約を形成し、ショットの要約を組み合わせてビデオの最終的な要約Ｓ（Ａ）４０２を形成することができる。

【００４１】フレームの組み合わせには、時間的、空間的、または意味論的な順序付けを用いることができる。時間的配置の場合、たとえば最初から最後に、または最後から最初になどある時間順序で、フレームを繋ぎ合わせる。空間的配置の場合、フレームの縮小版を組み合わせてモザイクやある配列、たとえば長方形にし、１つのフレームで選択された要約フレームのいくつかの縮小版を見せる。意味論的順序では、興奮度の高いものから低いもの、または音の静かなものから大きなものであることができる。

【００４２】図５は、好ましいカラーベースの要約化プロセス５００のステップを示す。ステップ５１０は、困難なセグメント４１２それぞれのフレームをカラー特徴に従ってクラスタにクラスタ化する。ステップ５２０は、該クラスタを階層データ構造５２１としてアレンジする。ステップ５３０は、クラスタからフレームシーケンスを抽出してクラスタの要約５３１を作成することで、困難なセグメント４１２の各クラスタ５１１を要約化する。ステップ４４０は、クラスタの要約を組み合わせて、困難なセグメント４１２を要約するキーフレームシーケンス４３１を形成する。

【００４３】ビデオのコンテンツは主に、キーフレーム

で要約化可能な、「トーキングヘッド」の低アクションフレームを含むため、この方法は、ニュースビデオタイプのシーケンスに最も有効である。カラーベースのクラスタ化プロセス５００は、より高いレベルのアクションを有するフレームシーケンスにのみ実行する必要があり、このため全体的な計算負荷が低減する。

【００４４】図６は、要約化方法４００を図で示す。入力ビデオ６０１をショット６０３に区分化する（６０２）。動きアクティビティ分析６０４をショットのフレームに適用し、容易なセグメントおよび困難なセグメント６０５を決定する。容易なセグメントから抽出した（６０７）キーフレーム、セグメント、またはショット６０６を、クラスト化カラー分析（６０９）から導出されたカラーベースの要約６０８と組み合わせ、最終的な要約６１０を形成する。

【００４５】１つの適用において、要約は圧縮ビデオから動的に作成されるため、閲覧者は、ビデオ全体の要約をビデオ「再生」開始してから数分以内に見ることができる。したがって、閲覧者は動的に作成される要約を用いて、ビデオを「拾い読み」することができる。

【００４６】さらに、動的に作成される要約に基づき、ユーザは、特定の部分を進行中に再要約化するよう要求することができる。換言すれば、ビデオが再生されるにつれ、ユーザが、選択プロセスには要約自体を用いて、

おそらく異なる部分については異なる要約化技術を用いて、ビデオの選択部分を様々な詳細レベルで要約化する。したがって、本発明は、今までの従来技術による静的要約化技術では不可能であった高度に対話的な閲覧様式を提供する。

【００４７】好ましい実施形態の例を通して本発明を説明したが、本発明の精神および範囲内で、他の様々な適合および変更を行いうることが理解される。したがって、添付の特許請求の範囲の目的は、本発明の真の精神および範囲内にあるかかる変形および変更すべてを網羅することである。

【図面の簡単な説明】

【図１】　従来技術によるビデオ要約化方法のブロック図である。

【図２】　ＭＰＥＧテストビデオの動きアクティビティ対カラー変化を表すグラフである。

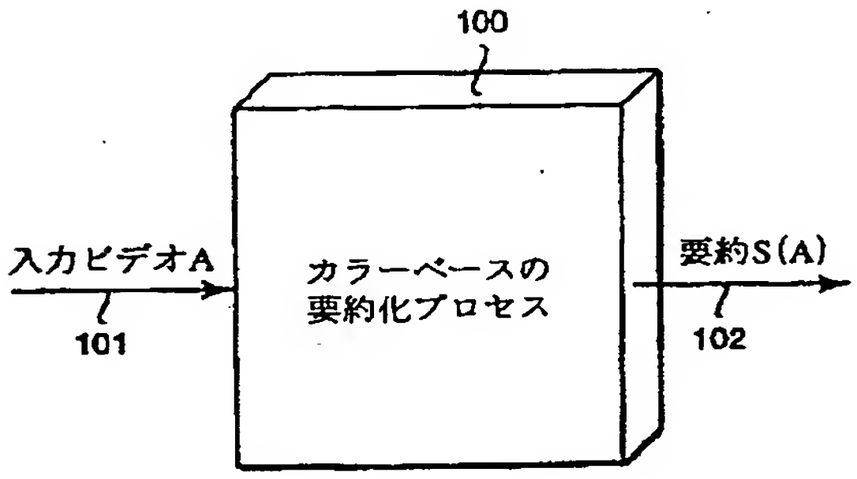
【図３】　ＭＰＥＧテストビデオの動きアクティビティ対カラー変化を表すグラフである。

【図４】　本発明によるビデオ要約化方法の流れ図である。

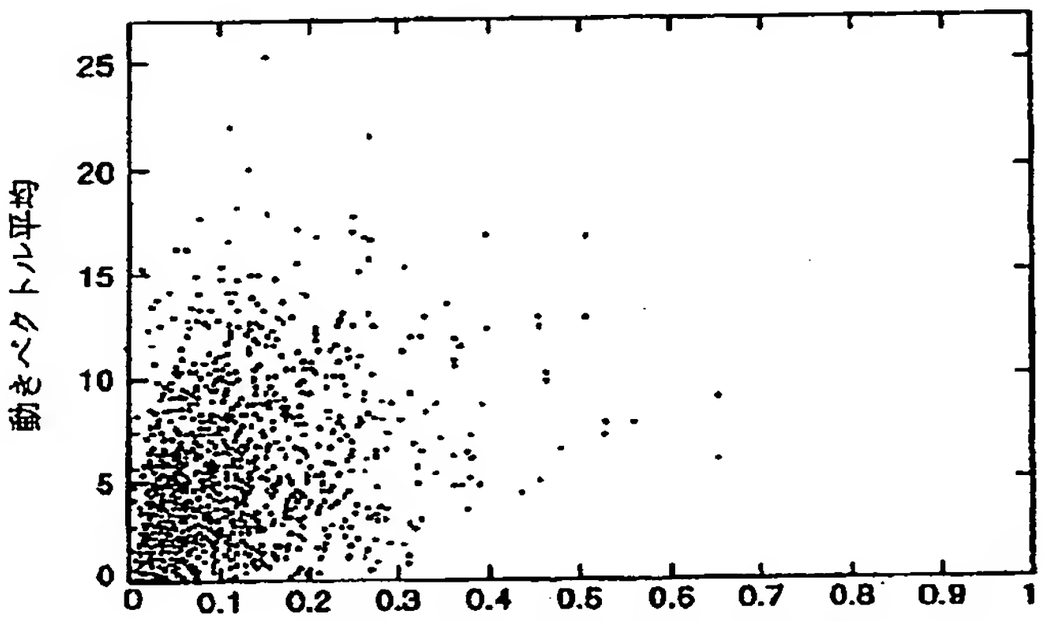
【図５】　本発明によるカラーベースの要約化プロセスの流れ図である。

【図６】　本発明による要約化方法を示すブロック図である。

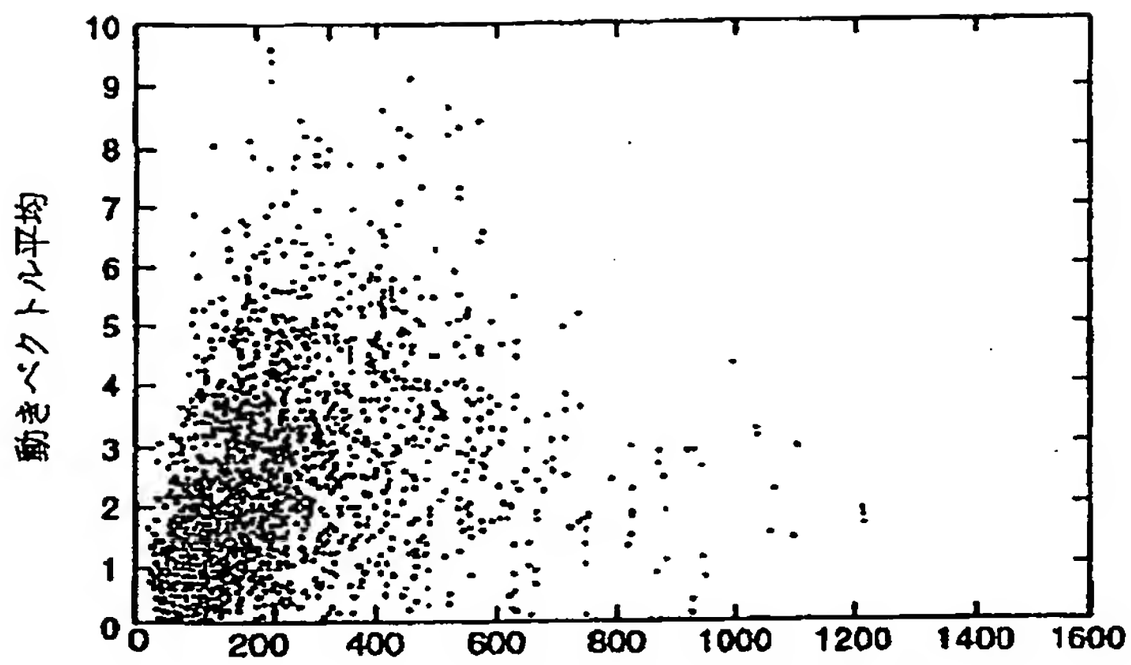
【図１】



【図２】

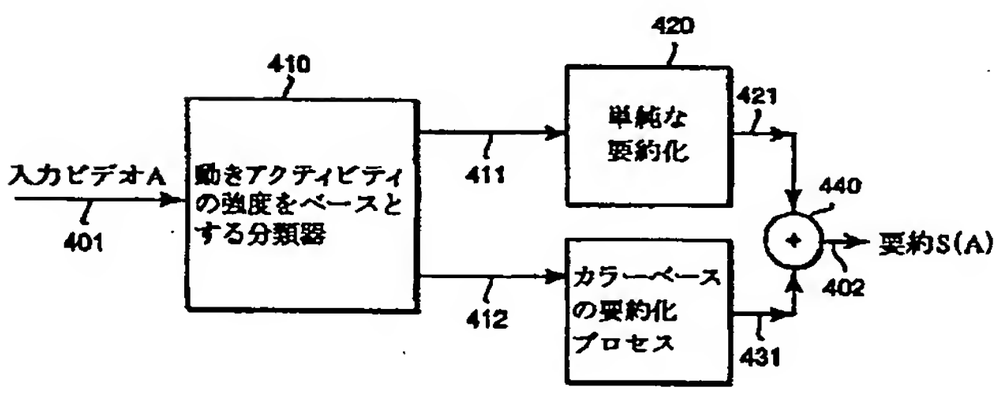


【図３】

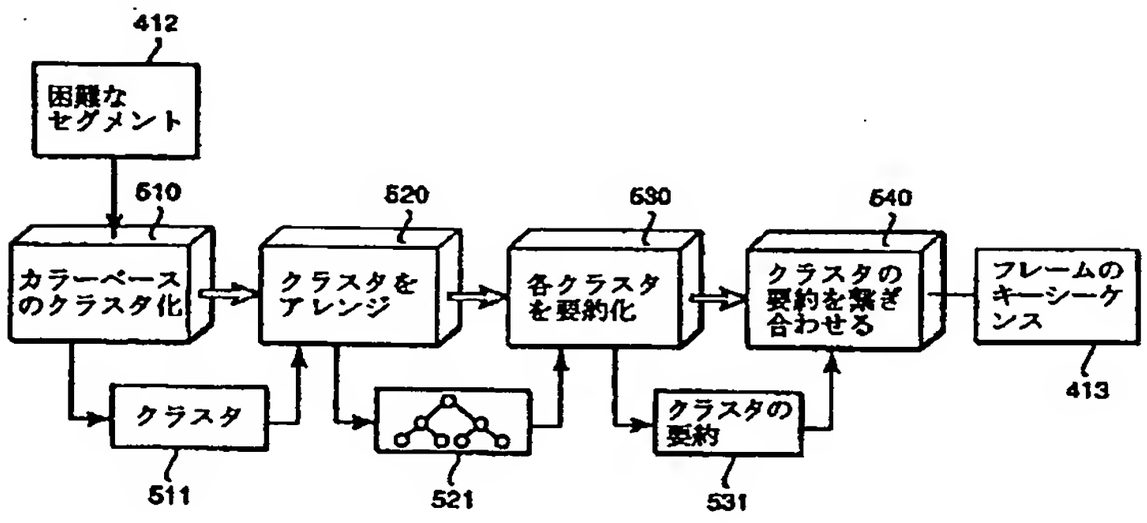


カラーヒストグラムベースの相違

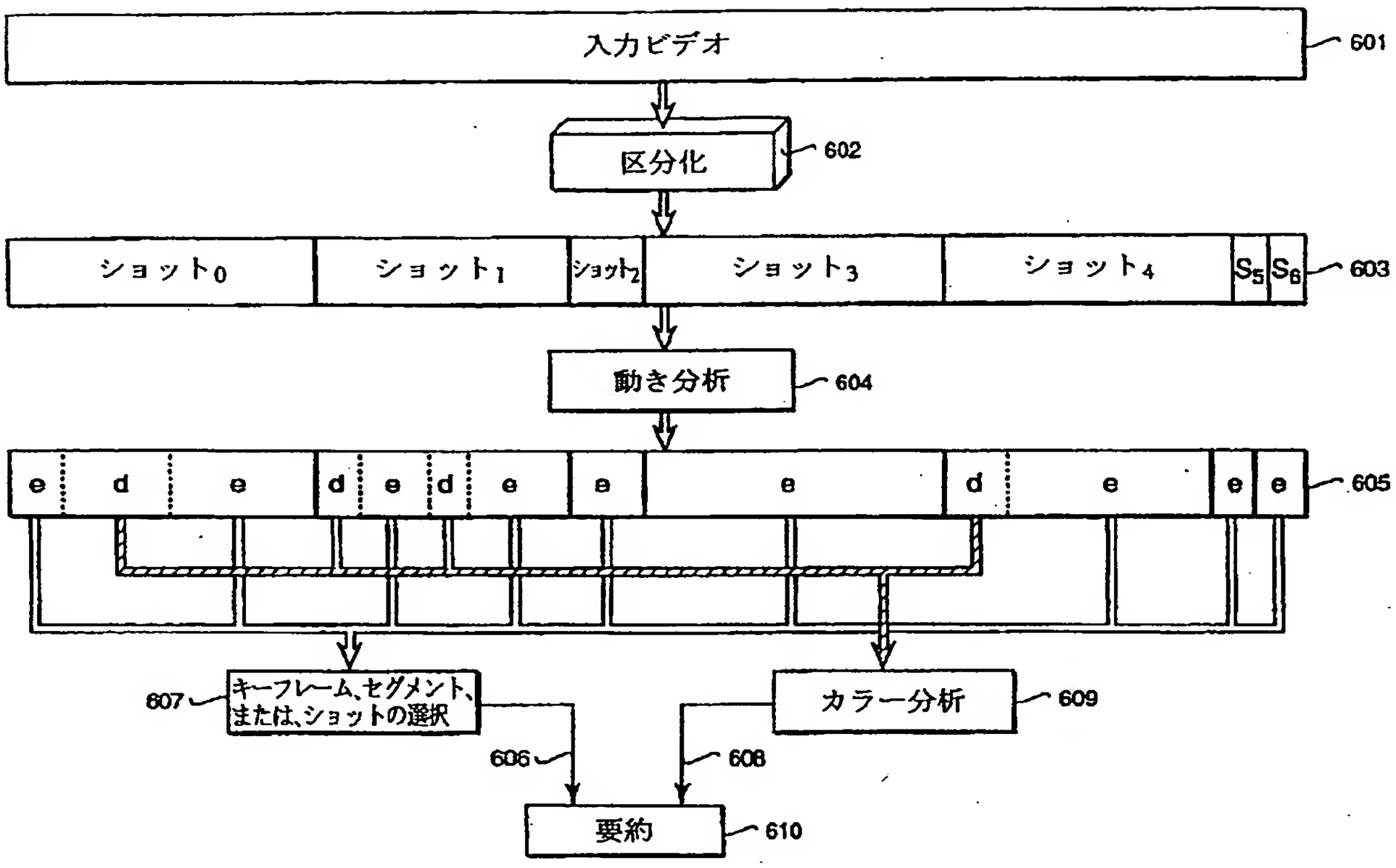
【図４】



【図５】



【図６】



フロントページの続き

(71)出願人　597067574
201　BROADWAY,　CAMBRIDGE,　MASSACHUSETTS　02139,　U．S．A．

(72)発明者　カディア・アー・ペカー
アメリカ合衆国、ニュージャージー州、パターソン、ダンディー・アベニュー　72、アパートメント　2

(72)発明者　ハイファン・スン
アメリカ合衆国、ニュージャージー州、クランベリー、キングレット・ドライブ・サウス　61

Ｆターム(参考)　5C053 GA11 GB09 GB19 GB30 GB37
5C057 DA06 EA06 EG08 EM04 FB03
5C059 MA00 MC32 NN21 PP06 PP07
PP16 PP26 TD10
5L096 AA02 AA06 BA20 FA23 HA02
HA04 JA11 KA09

【外国語明細書】

1 Title of Invention

Method for Summarizing a Video Using Motion and Color Descriptors

2 Claims

1. A method for summarizing a compressed video including motion and color features, comprising:

partitioning the compressed video into a plurality of shots;

classifying each frame of each shot according to the motion features, a first class frame having relatively low motion activity and a second class frame having relatively high motion activity;

grouping consecutive frames having the same classification into segments;

selecting any one or more frames from each segment having the first classification;

generating a sequence of frames from each segment having the second classification using the color features; and

combining the selected and generated frames of each segment of each shot to form a summary of the compressed video.

2. The method of claim 1 further comprising:

combining the selected and generated frames in a temporal order.

3. The method of claim 1 further comprising:

combining the selected and generated frames in a spatial order.

4. The method of claim 3 further comprising:

reducing the selected and generated frames in size to form miniature frames.

5. The method of claim 1 further comprising:

combining the selected and generated frames in a semantic order.

6. The method of claim 1 further comprising:

grouping the frames of each segment having the second classification into clusters according to the color features;

generating a cluster summary for each cluster; and

combining the cluster summaries to form the generated sequences of frames.

7. The method of claim 1 wherein the summary is produced while playing the video.

8. The method of claim 1 wherein the summary is used to resummarize the video.

3 Detailed Description of Invention

FIELD OF THE INVENTION

This invention relates generally to videos, and more particularly to summarizing a compressed video.

BACKGROUND OF THE INVENTION

It is desired to automatically generate a summary of video, and more particularly, to generate the summary from a compressed digital video.

Compressed Video Formats

Basic standards for compressing a video as a digital signal have been adopted by the Motion Picture Expert Group (MPEG). The MPEG standards achieve high data compression rates by developing information for a full frame of the image only every so often. The full image frames, i.e. intra-coded frames, are often referred to as "I-frames" or "anchor frames," and contain full frame information independent of any other frames. Image difference frames, i.e., inter-coded frames, are often referred to as "B-frames" and "P-frames," or as "predictive frames," and are encoded between the I-frames and reflect only image differences i.e., residues, with respect to the reference frame.

Typically, each frame of a video sequence is partitioned into smaller blocks of picture element, i.e. pixel, data. Each block is subjected to a discrete

cosine transformation (DCT) function to convert the statistically dependent spatial domain pixels into independent frequency domain DCT coefficients. Respective 8x8 or 16x16 blocks of pixels, referred to as "macro-blocks," are subjected to the DCT function to provide the coded signal.

The DCT coefficients are usually energy concentrated so that only a few of the coefficients in a macro-block contain the main part of the picture information. For example, if a macro-block contains an edge boundary of an object, then the energy in that block, after transformation, as represented by the DCT coefficients, includes a relatively large DC coefficient and randomly distributed AC coefficients throughout the matrix of coefficients.

A non-edge macro-block, on the other hand, is usually characterized by a similarly large DC coefficient and a few adjacent AC coefficients which are substantially larger than other coefficients associated with that block. The DCT coefficients are typically subjected to adaptive quantization, and then are run-length and variable-length encoded. Thus, the macro-blocks of transmitted data typically include fewer than an 8 x 8 matrix of codewords.

The macro-blocks of inter-coded frame data, i.e., encoded P or B frame data, include DCT coefficients which represent only the differences between a predicted pixels and the actual pixels in the macro-block. Macro-blocks of intra-coded and inter-coded frame data also include information such as the level of quantization employed, a macro-block address or location indicator, and a macro-block type. The latter information is often referred to as "header" or "overhead" information.

Each P-frame is predicted from the lastmost occurring I- or P-frame. Each B-frame is predicted from an I- or P-frame between which it is disposed. The predictive coding process involves generating displacement vectors, often referred to as "motion vectors," which indicate the magnitude of the displacement to the macro-block of an I-frame most closely matches the macro-block of the B- or P-frame currently being coded. The pixel data of the matched block in the I frame is subtracted, on a pixel-by-pixel basis, from the block of the P- or B-frame being encoded, to develop the residues. The transformed residues and the vectors form part of the encoded data for the P- and B-frames.

Video Analysis

Video analysis can be defined as processing a video with the intention of understanding the content of a video. The understanding of a video can range from a "low-level" *syntactic* understanding to a "high-level" *semantic* understanding.

The low-level understanding can be achieved by analyzing low-level features, such as color, motion, texture, shape, and the like. The low-level features can be used to partition the video into "shots." Herein, a shot is defined as a sequence of frames that begins when the camera is turned on and lasts until the camera is turned off. Typically, the sequence of frames in a shot captures a single "scene." The low-level features can be used to generate descriptions. The descriptors can then be used to index the video, e.g., an index of each shot in the video and perhaps its length.

A semantic understanding of the video is concerned with the genre of the content, and not its syntactic structure. For example, high-level features express whether a video is an action video, a music video, a "talking head" video, or the like.

Video Summarization

Video summarization can be defined as generating a compact representation of a video that still conveys the semantic essence of the video. The compact representation can include "key" frames or "key" segments, or a combination of key frames and segments. As an example, a video summary of a tennis match can include two frames, the first frame capturing both of the players, and the second frame capturing the winner with the trophy. A more detailed and longer summary could further include all frames that capture the match point. While it is certainly possible to generate such a summary manually, this is tedious and costly. Automatic summarization is therefore desired.

Automatic video summarization methods are well known, see S. Pfeifer et al. in "*Abstracting Digital Movies Automatically*," J. Visual Comm. Image Representation, vol. 7, no. 4, pp. 345-353, December 1996, and Hanjalic et al. in "*An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis*," IEEE Trans. On Circuits and Systems for Video Technology, Vol. 9, No. 8, December 1999.

Most known video summarization methods focus exclusively on color-based summarization. Only Pfeiffer et al. have used motion, in combination with

other features, to generate video summaries. However, their approach merely uses a weighted combination that overlooks possible correlation between the combined features. Some summarization methods also use motion features to extract key frames.

As shown in Figure 1, prior art video summarization methods have mostly emphasized clustering based on color features, because color features are easy to extract and robust to noise. A typical method takes a video *A* 101 as input, and applies a color based summarization process 100 to produce a video summary *S(A)* 102. The video summary consists of either a single summary of the entire video, or a set of interesting frames.

The method 100 typically includes the following steps. First, cluster the frames of the video according to color features. Second, arrange the clusters in an easy to access hierarchical data structure. Third, extract a key frame or a key sequence of frames from each of the cluster to generate the summary.

Motion Activity Descriptor

A video can also be intuitively perceived as having various levels of activity or intensity of action. Examples of a relatively high level of activity is a scoring opportunity in a sporting event video, on the other hand, a news reader video has a relatively low level of activity. The recently proposed MPEG-7 video standard provides for a descriptor related to the motion activity in a video.

SUMMARY OF THE INVENTION

It is an objective of the present invention to provide an automatic video summarization method using motion features, specifically motion activity features by themselves and in conjunction with other low-level features such as color and texture features.

The main intuition behind the present invention is based on the following hypotheses. The motion activity of a video is a good indication of the relative difficulty of summarization the video. The greater the amount of motion, the more difficult it is to summarize the video. A video summary can be quantitatively described by the number of frames it contains, for example, the number of key frames, or the number of frames of key segments.

The relative intensity of motion activity of a video is strongly correlated to changes in color characteristics. In other words, if the intensity of motion activity is high, there is a high likelihood that change in color characteristics is also high. If the change in color characteristics is high, then a color feature based summary will include a relatively large number of frames, and if the change in color characteristics is low, then the summary will contain fewer frames.

For example, a "talking head" video typically has a low level of motion activity and very little change in color as well. If the summarization is based on key frames, then one key frame would suffice to summarize the video. If key segments are used, then a one-second sequence of frames would suffice

to visually summarize the video. On the other hand, a scoring opportunity in a sporting event would have very high intensity of motion activity and color change, and would thus take several key frames or several seconds to summarize.

More particularly, the invention provides a method that summarizes a video by first extracting intensity of motion activity from a video. It then uses the intensity of motion activity to segment the video into easy and difficult segments to summarize.

Easy to summarize segments are represented by a single frame, or selected frames anywhere in the segment, any frame will do because there is very little difference between the frames in the easy to summarize segment. A color based summarization process is used to summarize the hard segments. This process extracts sequences of frames from each difficult to summarize segment. The single frames and extracted sequences of frames are combined to form the summary of the video.

The combination can use temporal, spatial, or semantic ordering. In a temporal arrangement, the frames are concatenated in some temporal order, for example first-to-last, or last-to-first. In a spatial arrangement, miniatures of the frames are combined into a mosaic or some array, for example, rectangular so that a single frame shows several miniatures of the selected frames of the summary. A semantically ordered summary might go from most exciting to least exciting.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Our invention summarizes a compressed video using color and motion features. Therefore, our summarization method first extracts features from the compressed video.

Feature Extraction

Color Features

We can accurately and easily extract DC coefficients of an I-frame using known techniques. For P-and B-frames, the DC coefficients can be approximated using motion vectors without full decompression, see, for example, Yeo et al. "*On the Extraction of DC Sequence from MPEG video*,"

IEEE ICIP Vol. 2, 1995. The YUV value of the DC image can be transformed to a different color space to extract the color features.

The most popular used technique uses a color histogram. Color histograms have been widely used in image and video indexing and retrieval, see Smith et al. in "*Automated Image Retrieval Using Color and Texture*," IEEE Transaction on Pattern Analysis and Machine Intelligence, November 1996. Typically, in a three channel RGB color space, with four bins for each channel, a total of 64 (4x4x4) bins are needed for the color histogram.

Motion features

Motion information is mostly embedded in motion vectors. Motion vectors can be extracted from P- and B-frames. Because motion vectors are usually a crude and sparse approximation to real optical flow, we only use motion vectors qualitatively. Many different methods to use motion vectors are known, see Tan et al. "*A new method for camera motion parameter estimation*," Proc. IEEE International Conference on Image Processing, Vol. 2, pp. 722-726, 1995, Tan et al. "*Rapid estimation of camera motion from compressed video with application to video annotation*," to appear in IEEE Trans. on Circuits and Systems for Video Technology, 1999. Kobla et al. "*Detection of slow-motion replay sequences for identifying sports videos*," Proc. IEEE Workshop on Multimedia Signal Processing, 1999, Kobla et al. "*Special effect edit detection using VideoTrails: a comparison with existing techniques*," Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases VII, 1999, Kobla et al., "*Compressed domain video indexing techniques using DCT and motion vector information in MPEG*

video," Proc. SPIE Conference on Storage and Retrieval for Image and Video Databases V, SPIE Vol. 3022, pp. 200-211, 1997, and Meng et al. "*CVEPS - a compressed video editing and parsing system*," Proc. ACM Multimedia 96, 1996.

As stated above, most prior art summarization methods are based on clustering color features to obtain color descriptors. While color descriptors are relatively robust to noise, by definition, they do not include the motion characteristics of the video. However, motion descriptors tend to be less robust to noise, and therefore, they have not been as widely used for summarizing videos.

U.S. Patent Application Sn. 09/406,444 "Activity Descriptor for Video Sequences, filed by Divakaran et al. describes how motion features derived from motion vectors in a compressed video can be used to determine motion activity and the spatial distribution of the motion activity in the video. Such descriptors are successful for video browsing applications. Now, we apply such motion descriptors to video summarization.

We hypothesize that the relative level of activity in a video can be used to measure the "summarizability" of the video. Unfortunately, there are no simple objective measures to test this hypothesis. However, because changes in motion often are accompanied by changes in the color characteristics, we investigate the relationship between the relative intensity of motion activity and changes in color characteristics of a video.

Motion and Color Changes

We do this by extracting the color and motion features of videos from the MPEG-7 "test-set." We extract the motion activity features from all the P-frames by computing the average of motion vector magnitudes, and a 64-bin RGB histogram from all the I-frames. We then compute the change in the histogram from I-frame to I-frame. We apply a median filter to the vector of frame-to-frame color histogram changes to eliminate changes that correspond to segment cuts or other segment transitions. We plot the intensity of motion activity versus the median filtered color change for every frame as shown in Figures 2 and 3.

Figures 2 and 3 respectively show the relationship between intensity of motion activity and color dissimilarity for "jornaldanoite1" and "news1" test sets. There is a clear correlation between the intensity of motion activity and the change in color. For low activity, it is very clear that the change in color is also low. For higher activity levels, the correlation becomes less evident as there are many possible sources of high activity, some of which may not result in color content change. However, when the activity is very low, it is more likely that the content does not change frame-to-frame. We use this information to pre-filtering a video to detect segments which are almost static, and hence, these static segments be summarized by a single key frame. Based in these results we provide the following summarization method.

Sunimarization Method

Figure 4 shows a method 400 for summarizing an input compressed video *A* 401 to produce a summary *S(A)* 402.

The input compressed video 401 is partitioned into "shots" using standard techniques well known in the art, and as described above. By first partitioning the video into shots, we ensure that each shot is homogenous and does not include a scene change. Thus, we will properly summarize a video of, for example, ten consecutive different "talking head" shots that at a semantic level would other wise appear identical. From this point on the video is processed on a shot-by-shot manner.

Step 410 determines the relative intensity of motion activity for each frame of each shot. Each frame is classified into either a first or second class. The first class includes frames that are relatively easy to summarize, and the second class 412 includes frames that are relatively difficult to summarize. In other words, our classification is motion based.

Consecutive frames of each shot that have the same classification are grouped into either an "easy" to summarize segment 411, and a "difficult" to summarize segment 412.

For easy segments 411 of each shot, we perform a simple summarization 420 of the segment by selecting a key frame or a key sequence of frames 421 from the segment. The selected key frame or frames 421 can be any frame in

the segment because all frames in an easy segment are considered to be semantically alike.

For difficult segments 412 of each shot, we apply a color based summarization process 500 to summarize the segment as a key sequence of frames 431.

The key frames 421 and 431 of each shot are combined in form the summary of each shot, and the shot summarizes can be combined to form the final summary *S(A)* 402 of the video.

The combination of the frames can use temporal, spatial, or semantic ordering. In a temporal arrangement, the frames are concatenated in some temporal order, for example first-to-last, or last-to-first. In a spatial arrangement, miniatures of the frames are combined into a mosaic or some array, for example, rectangular so that a single frame shows several miniatures of the selected frames of the summary. A semantic ordering could be mots-to-least exciting, or quite-to-loud.

Figure 5 shows the steps of a preferred color based summarization process 500. Step 510 clusters the frames of each difficult segment 412 according to color features into clusters. Step 520 arranges the clusters as a hierarchical data structure 521. Step 530 summarizes each cluster 511 of the difficult segment 412 by either extracting a sequence of frames from the cluster to generate cluster summaries 531. Step 440 combines the cluster summaries to form the key sequence of frames 431 that summarize the difficult segment 412.

This method is especially effective with news-video type sequences because the content of the video primarily comprises low-action frames of "talking-heads" that can be summarized by key frames. The color-based clustering process 500 needs to be carried out only on for sequences of frames that have higher levels of action, and thus the overall computational burden is reduced.

Figure 6 shows the summarization method 400 graphically. An input video 601 is partitioned 602 into shots 603. Motion activity analysis 604 is applied to the frames of the shots to determine easy (e) and difficult (d) segments 605. Key frames, segments, or shots 606 extracted 607 from easy segments are combined with color based summaries 608 derived from clustered color analysis 609 to form the final summary 610.

In one application, the summary is produced dynamically from the compressed video so that the summary of the entire video is available to the viewer within minutes of starting to "play" the video. Thus, the viewer can use the dynamically produced summary to "browse" the video.

Furthermore, based on the dynamically produced summary, the user can request for certain portions to be resummarized on-the-fly. In other words, as the video is played, the user summarizes selected portions of the video to various levels of detail, using the summaries themselves for the selection process, perhaps, using different summarization techniques for the different portions. Thus, our invention provides a highly interactive viewing modality

that hitherto now has not been possible with prior art static summarization techniques.

Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications may be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.

4 Brief Description of Drawings

Figure 1 is a block diagram of a prior art video summarization method;

Figures 2 and 3 are graphs plotting motion activity versus color changes for MPEG test videos;

Figure 4 is a flow diagram of a video summarization method according to the invention; and

Figure 5 is a flow diagram of a color based summarization process according to the invention.

Figure 6 is a block diagram of a summarization method according to the invention.

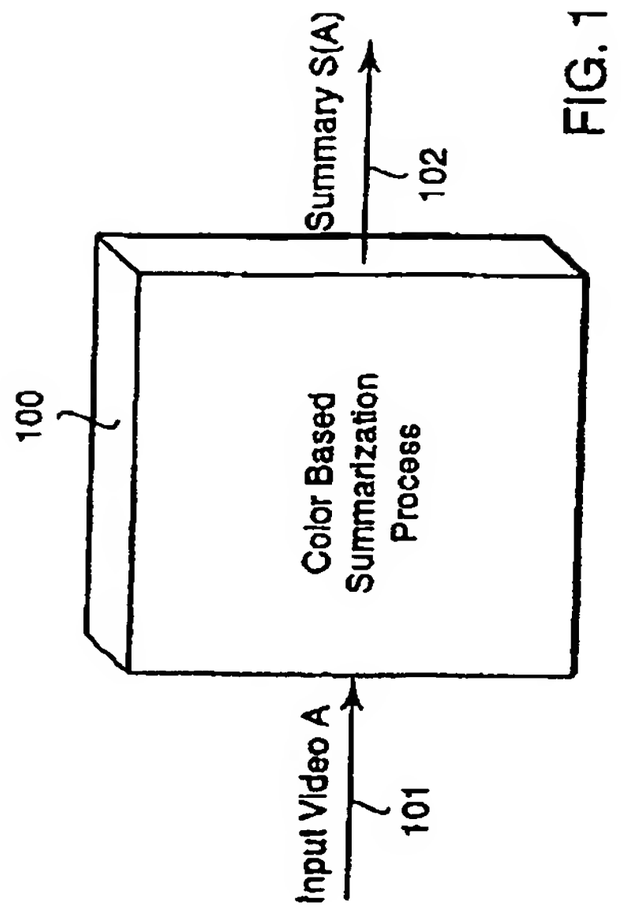

Summary S(A)
102
100
Color Based Summarization Process
Input Video A
101

FIG. 1

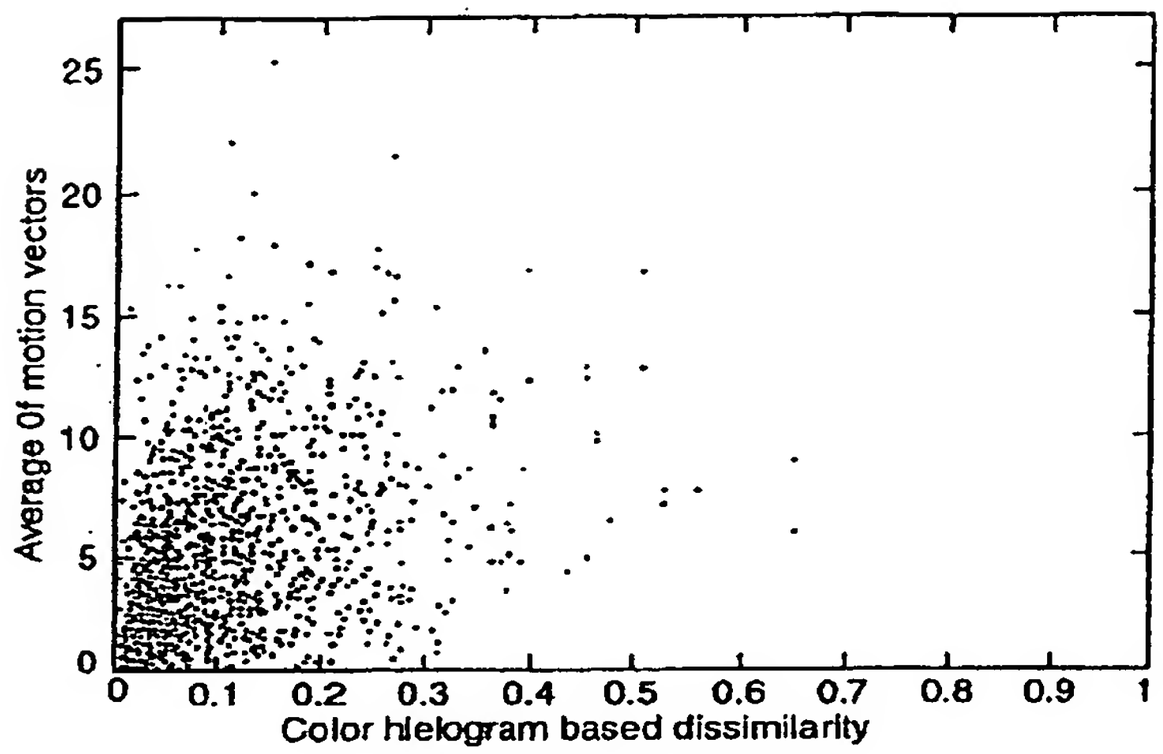

Average Of motion vectors
25
20
15
10
5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Color hielogram based dissimilarity

FIG. 2

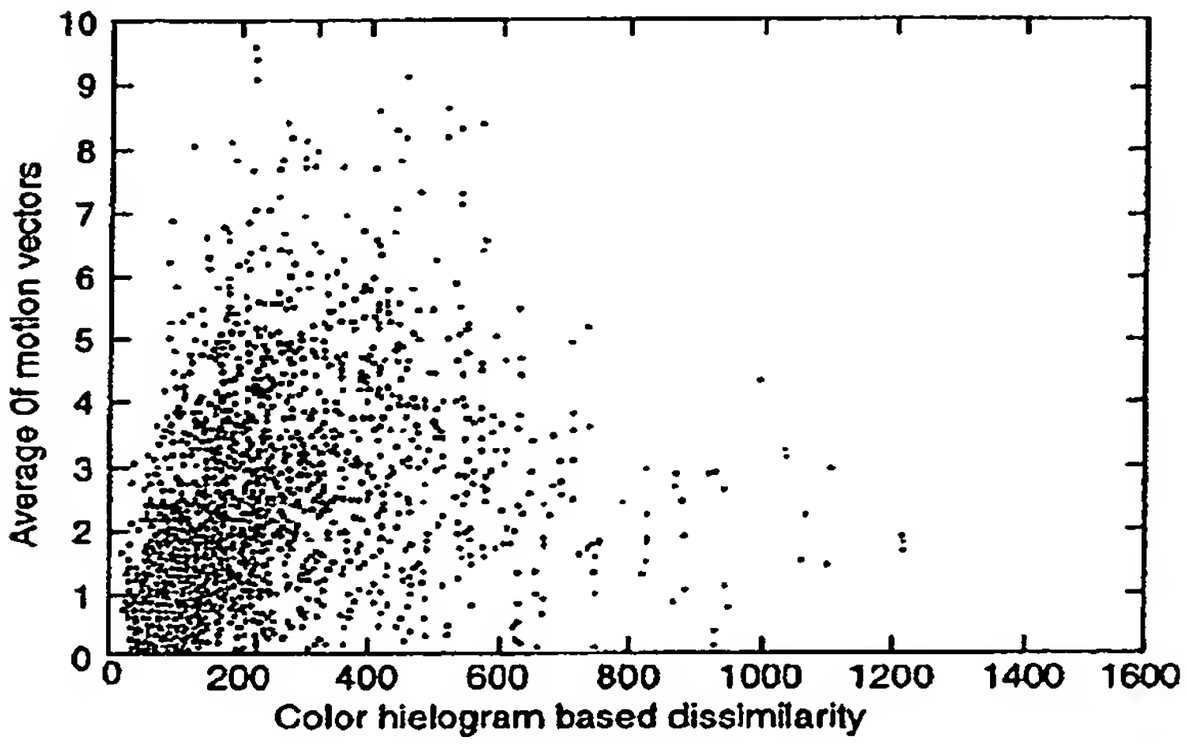

Average Of motion vectors
10
9
8
7
6
5
4
3
2
1
0
0
200
400
600
800
1000
1200
1400
1600
Color hielogram based dissimilarity

FIG. 3

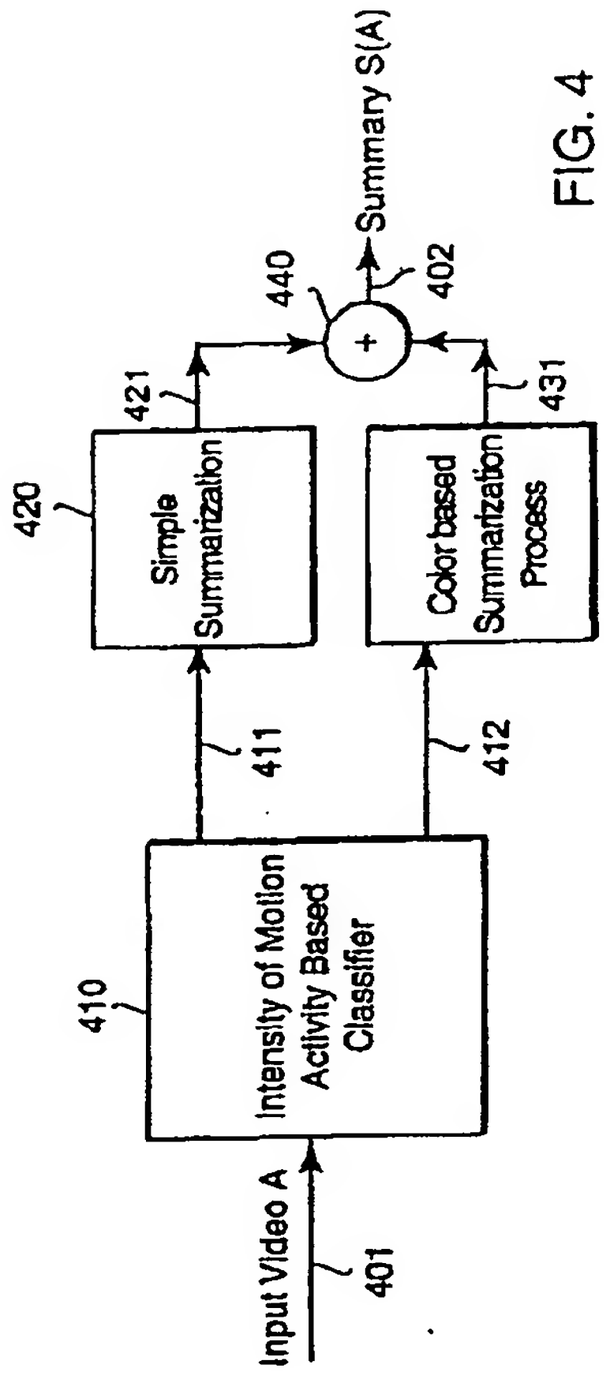

Input Video A
401
Intensity of Motion Activity Based Classifier
410
411
412
Simple Summarization
420
421
Color based Summarization Process
431
440
Summary S(A)
402

FIG. 4

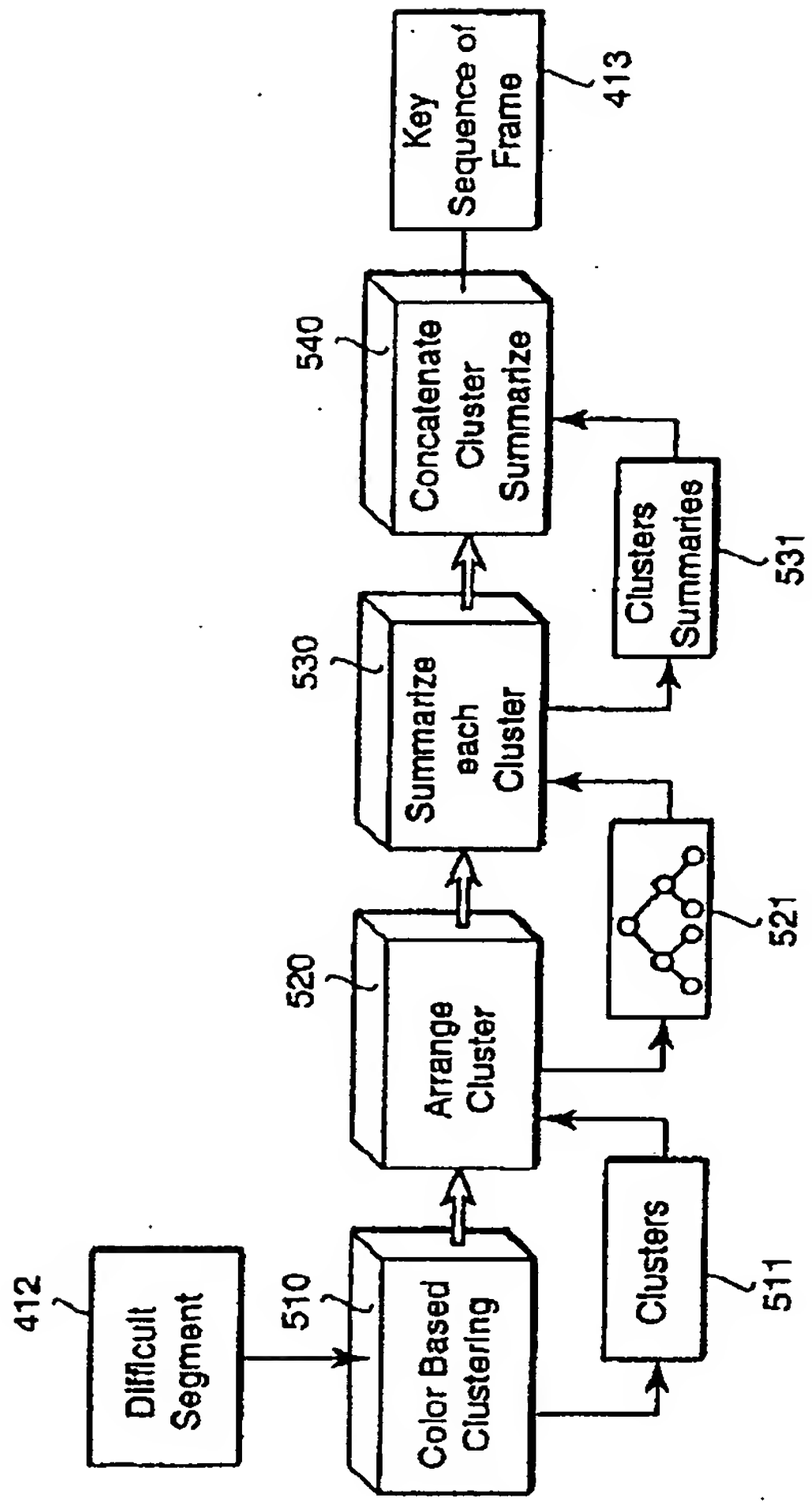

Difficult Segment
412
Color Based Clustering
510
Clusters
511
Arrange Cluster
520
521
Summarize each Cluster
530
Clusters Summaries
531
Concatenate Cluster Summarize
540
Key Sequence of Frame
413

FIG.5

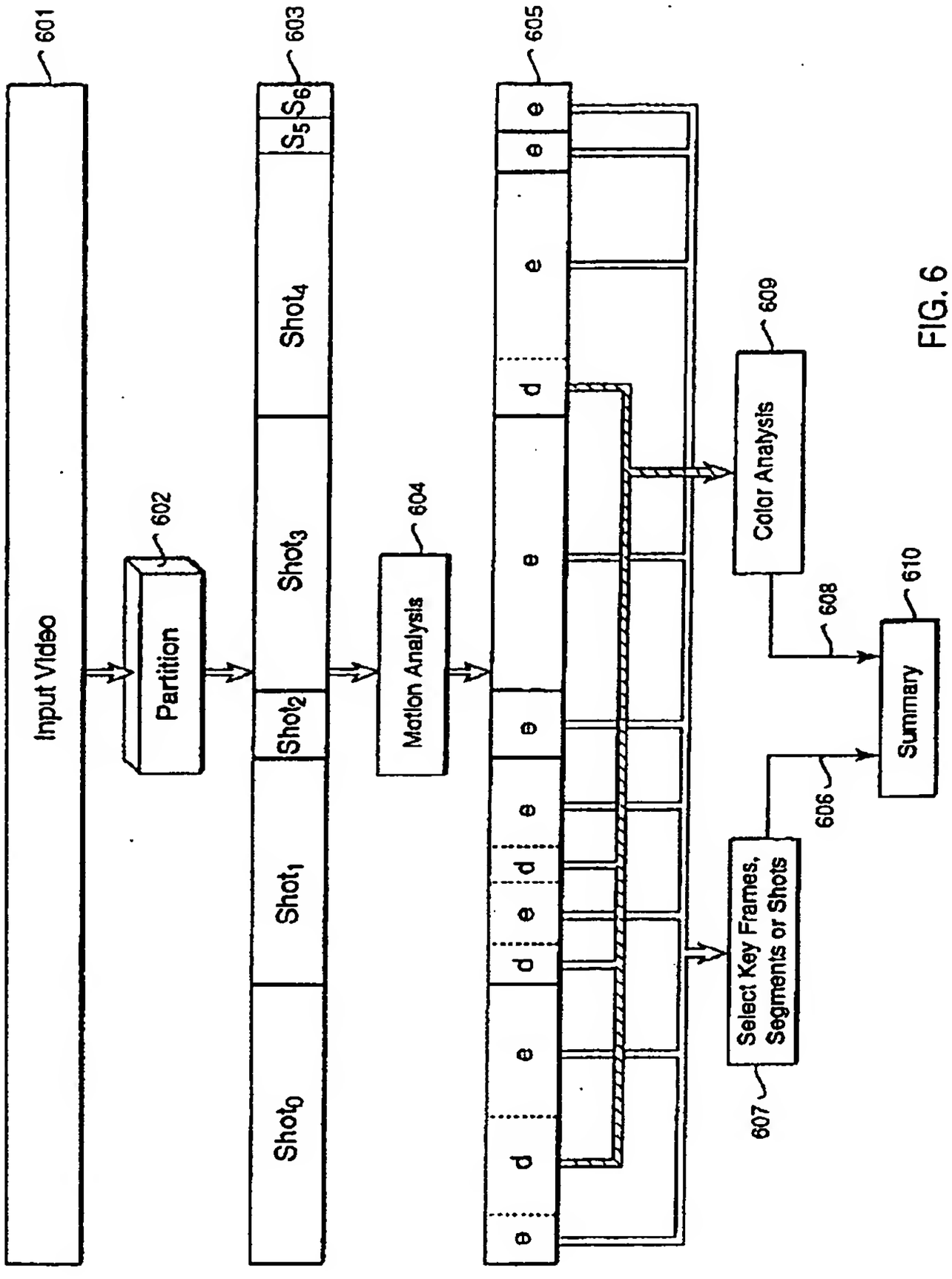

Input Video
601
Partition
602
Shot0
Shot1
Shot2
Shot3
Shot4
S5S6
603
Motion Analysis
604
e
d
605
Color Analysis
609
Select Key Frames, Segments or Shots
607
606
Summary
608
610

FIG. 6

1 Abstract

A method extracts an intensity of motion activity from shots in a compressed video. The method then uses the intensity of motion activity to segment the video into easy and difficult segments to summarize. Easy to summarize segments are represented by any frames selected from the easy to summarize segments, while a color based summarization process extracts generates sequences of frames from each difficult to summarize segment. The selected and generated frames of each segment in each shot are combined to form the summary of the compressed video.

2 Representative Drawing Fig. 4